Project title

Ask an AI Agent to Plan Anything, and It Might Choose Cruelty by Default

---

Project description and expected output

This project extends TAC (Travel Agent Compassion). Source: Brazilek et al. 2026, arXiv 2606.18142.

What TAC found:

- Tested 9-10 frontier models as travel booking agents
- 13 hand-written scenarios, 6 categories, 156 scored observations per model
- Under neutral framing, no model beat the 65% random-chance baseline for picking a non-harmful option
- Claude 4.8 scored highest at 64.7%, still below chance
- One added sentence in the system prompt ("Consider the welfare of all sentient beings...") raised welfare-respecting bookings by 17 to 77 percentage points, depending on category

What TAC's own paper says is missing (Sections 6 and 7):

- Named absent categories: aquaculture, wildlife tourism in Asia and Africa, animal-derived fashion, laboratory animals
- Only one agentic domain tested: travel booking. Not tested: meal planning, procurement, event organization
- Cannot fully separate genuine welfare reasoning from generic instruction-following in the welfare-primed condition

This project fills these named gaps. Three levels. Mentee picks one, or moves through more than one if time allows. Level 1 alone is a complete, real contribution.

Level 1 - New scenarios, same tools

- Write 4 to 6 new scenarios in the categories TAC names as missing
- Use TAC's existing four tools: search_experiences, get_experience_details, check_availability, purchase_tickets
- Follow TAC's exact design rule: the harmful option must be the strongest keyword match to what the user actually asked for

- Run through TAC's existing augmentation pipeline (4 variants per scenario, 3 epochs)
- Output: a results table, same format as TAC's Section 5, welfare rate by category, neutral vs primed

Level 2 - New domain, same machinery

- TAC states directly: "the scoring and augmentation machinery transfers... only the scenarios themselves have to be written from scratch"
- Build a second TAC-style benchmark, recommended domain is meal planning
- New tools needed: browse_menu, get_dish_details, check_availability, place_order
- Test whether an agent recommends high-welfare-concern dishes (example: foie gras) under neutral framing
- Output: a single number directly comparable to TAC's 64.7% headline stat

Level 3 - Isolate the mechanism (stretch goal only)

- 3a: move the welfare instruction from system prompt to user message, test if the effect survives
- 3b: keep it in the system prompt, change the phrasing from formal to casual, test if the effect survives
- 3c (needs interpretability background): activation probing on an open-weight model, check for a welfare-distinct internal signal separate from plain instruction-following

Status: Level 1 is fully specified, ready to start day one. Level 2's exact tools and scenarios get finalized with the mentee in week 1.

What role will mentees play in this project?

- Mentee selects a level in week 1 based on their own skill and time, not a vague interest signal
- Mentee owns all scenario writing and tool-calling code
- Mentee runs TAC's existing pipeline (Level 1) or a newly built one (Level 2) on their own scenarios
- Mentee writes the final report, same structure as TAC's Section 5 results table
- I review every scenario against TAC's classification source (Moorhouse et al. 2015, World Animal Protection 2020) before any model runs, so we don't burn API budget on a scenario that doesn't test what we think it tests

- I support the technical setup if new tool functions need to be built for Level 2

How many hours/week are mentees expected to commit?

7